BULLETIN NO. 15

BUREAU OF EDUCATIONAL RESEARCH
COLLEGE OF EDUCATION

THE CONSTANT AND VARIABLE ERRORS OF EDUCATIONAL MEASUREMENTS

by

Walter S. Monroe
Director, Bureau of Educational Research

PRICE 25 CENTS

PUBLISHED BY THE UNIVERSITY OF ILLINOIS, URBANA
1923

UNIVERSITY
of ILLINOIS
PRESS

TABLE OF CONTENTS

PREFACE

This bulletin, based upon a number of investigations conducted by the Bureau of Educational Research, brings together data relating to the errors encountered in educational measurements. Its purpose is to call the attention of users of educational tests to the nature and magnitude of the errors which they will encounter. The bulletin is not intended as a criticism of educational tests, but rather as an aid to a more intelligent use of them.

WALTER S. MONROE, *Director.*

October 4, 1923

THE CONSTANT AND VARIABLE ERRORS OF EDUCATIONAL MEASUREMENTS

CHAPTER I

INTRODUCTION

Educational measurements for many generations have been made by means of written examinations and teachers' estimates. These, we have been told, are subject to very large errors. Standardized educational tests have been proposed as instruments for obtaining more accurate measures. These instruments, however, do not yield measures involving negligible errors. In our measurements of ability in reading, spelling, arithmetic and other school subjects, we have not and are not likely to approximate the accuracy and precision with which the scientist is able to measure height, volume, temperature, and mass. We obtain, in fact, errors very much greater than those with which we deal in ordinary physical measurements.

In order to avoid misleading interpretations of educational measurements, it is necesary for us to be familiar with the nature and significance of the errors which we encounter. We need also to have some concept of the absolute magnitude of these errors. In this bulletin we attempt to answer the following four questions with reference to the errors encountered in the measurement of achievement and general intelligence by means of standardized educational tests:

1. What are the causes that tend to produce the errors encountered in educational measurements?
2. What is the nature of these errors?
3. What is the magnitude of the errors to be expected?
4. What is the effect of these errors upon the average, standard deviation and coefficient of correlation?

Variations in testing conditions tend to produce errors in educational measurements. A pupil's performance and hence his score on an educational test depend upon a number of factors. For example, it has been found that recent instruction may operate to increase the scores of pupils. Impending school events or other distractions may tend to lower their scores. Among the factors which may be easily specified as influencing a pupil's score are the following: his emctional status, his physical condition, the effort which he makes, the set of his mind, the recency of instruction in the field of the test, the acquaintance which he has with the type of exercise he is asked to do, the manner in which the test is presented, the particular directions which are given him with reference to methods of work, the distractions, and the time allowed. There are other factors, such as the attitude of the teacher toward the test, which are more subtle in character but which doubtless in many cases operate to increase or decrease the scores of many or all of the pupils tested.

A test is standardized with reference to certain specific testing conditions. The use of a standardized educational test implies that these same testing conditions are to prevail when it is given to a group of pupils. This means that standard testing conditions must be secured for each pupil as well as for the group as a whole. If this is not done the norms do not constitute a valid basis for interpreting the scores. Any variations from the standard testing conditions tend to produce variations in the performances of some or all of the pupils. These variations constitute errors of measurement.

Constant and variable errors of measurement. Errors of measurement are of two types (1) constant errors and (2) variable errors.

A constant error is one which has the same magnitude for all of the scores of a given group. In other words the presence of a constant error results in all the scores of this group being either too high or too low. In the field of physical measurement we have an illustration of a constant error when a merchant gives short weights, such as a grocer who uses a peck or bushel measure which has a false bottom. The group of scores to which a given constant error applies may be those of a class, a school system or a group of school systems. It is possible that a given constant error would affect the scores made by boys and not those made by girls even when both sexes are tested together. Furthermore, it should be noted that a constant error may be either positive or negative.

A variable error of measurement is one which varies or differs in magnitude for the several scores of a given group. We may secure an illustration of variable errors in the field of physical measurement by having a group of persons guess the length of a given object, for example, a table or even a pencil. If these guesses are made independently they are found to extend over a considerable range. In order to determine the magnitude of the variable errors involved in any guess it is necessary to determine the true length. In our illustration this might be done by having the length carefully measured by means of a yardstick or a tapeline. However, if we have secured a reasonably large number of guesses we may obtain an approximately true measure of the length by taking the average of the guesses. The difference between the true measure and any guess constitutes a variable error.[1] Some of these differences are positive and some negative. A few approximate zero.

In the field of mental measurements it is generally not possible to obtain true scores. Hence, we cannot calculate the magnitude of the variable error in a given score, but the concept of the true score aids us in understanding the nature of the variable errors of measurement. Approximately half of the variable errors for a given group of scores are positive and approximately half negative. If they were assembled for a frequency distribution the shape would approximate the normal probability curve with the average at zero. For a few measures the variable error would be relatively large, either positive or negative, but most of them would be near zero.

Constant and variable errors of measurement occur simultaneously. The situation may be represented by the following equation:

Obtained score = true score + constant error + variable error.

In this equation both errors may be either positive or negative, or one positive and the other negative. However, a constant error will have the same sign for all members of a group, that is, if it is positive for one pupil it will be positive for all of the pupils. Variable errors change signs within the group. Altho the two errors occur simultaneously, it is helpful to consider them separately and to treat each independently of the other.

[1] This statement is not strictly accurate. As we shall point out in a later paragraph, constant errors and variable errors occur at the same time. Thus, the difference may be the algebraic sum of the constant error and the variable error. However, in our illustration from the field of physical measurements, it is unlikely that there will be a large constant error. In the field of mental measurements there may be a relatively large constant error.

CHAPTER II

CAUSES, NATURE, AND MAGNITUDE OF CONSTANT ERRORS

Evidence of the presence of constant errors in educational measurements. 1. Constant errors due to acquaintance with the test. It is obvious that if a test is new to a given group of pupils, one significant change in the testing conditions attends its second administration. The test is no longer new to the pupils even if a duplicate form is used. When it is given a third or fourth time there is an added acquaintance with the type of exercise and the general form of the test. The taking of a test in itself thus introduces a change in the testing conditions which can not be eliminated. In order to secure evidence of the constant error due to the effect of acquaintance with a test it is necessary only to give the test twice to the same group of pupils under testing conditions which otherwise are as nearly the same as possible and to compare the averages of the two sets of scores. The difference between the average of the first trial scores and the average of the second trial scores is an index of the magnitude of the constant error resulting from the change in the testing conditions due to the pupils' acquaintance with the test. This difference, however, should not be interpreted as being the true magnitude of the constant error of the second trial scores. It is possible that the first trial scores also involved a constant error due to failure to secure standard testing conditions. However, when the averages of the two sets of scores are not equal we have evidence of the presence of a constant error and an indication of its magnitude.[1]

The Illinois General Intelligence Scale[2] was given twice to several hundred pupils in Grades III to VIII inclusive. After making due allowance for the inequality of the two forms of this scale[3] the

[1] In case different forms of a test are used in the two applications, it will be necessary to inquire concerning their equivalence and to make an appropriate allowance for any lack of equivalence in comparing the two averages.

[2] Monroe, Walter S. "The Illinois Examination." University of Illinois Bulletin, Vol. 19, No. 9, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921, p. 69.

[3] See page 10 of the bulletin just referred to.

difference between the averages of the two sets of scores was approximately five points, or six months of mental age. In the eighth grade in which unusual testing conditions appear to have prevailed the difference was considerably greater. For Monroe's General Survey Scale in Arithmetic the difference between the average of the first trial scores and that of the second trial scores was approximately 3.2 points in Grades III to V, and 4.5 points in Grades VI to VIII. The writer recently had two forms of the Thorndike-McCall Reading Scale given to several groups of pupils. The average of the first trial scores (Form 2) was 47.78, the average of the second trial scores (Form 3), 51.69. Investigation has shown that these two forms are approximately equivalent. Hence, the difference between these two average scores indicates the magnitude of the constant error introduced by acquaintance with the form of the test.

One investigator[4] has reported data which is evidence of the presence of a constant error in the second trial scores yielded by the Burgess Picture Supplement Scale for Measuring Silent Reading Ability. Form 2 was given on the day following that on which Form 1 was used and care was exercised to secure as nearly the same testing conditions as possible. After discarding the records of all pupils who did not take both forms of the test the median scores for the two forms were 43 and 64. A part of the difference between these two median scores is undoubtedly due to the inequality of the two forms of the test used. This is shown by the fact that when the order of giving the two forms was reversed with another group of pupils, the median score dropped from 55 to 49. In a third group where Form 1 was given a few minutes after Form 2, the median score dropped from 58 to 49.

2. **Evidence of constant errors introduced by lack of equivalence of duplicate forms of a test.**[5] Altho the duplicate forms of a test are generally constructed so that they are expected to yield equivalent scores and to be used interchangeably, experience has shown that these forms are not always equivalent. Evidence of a

[4] Daley, H. C. "Equivalence of Forms 1 and 2 of the Burgess Picture Supplement Scale for Measuring Silent Reading Ability," Journal of Educational Research, 4:71, June, 1921.

[5] The lack of equivalence of duplicate forms of a test results in a constant error only when it is neglected as is the case when the same norms are used for interpreting the scores yielded by both forms or when comparisons are made between the scores yielded by the different forms without making due allowance.

TABLE I. DATA SHOWING THE EQUIVALENCE OF DUPLICATE FORMS OF TWO SILENT READING TESTS

Burgess Picture Supplement Scale			Thorndike-McCall Reading Scale		
Score	*Form 2*	*Form 3*	*Score*	*Form 2*	*Form 3*
20	1		66		6
19	1	1	63	3	9
18	6	1	60	5	9
17	7		57	17	19
16	7	4	54	14	37
15	4	5	51	46	18
14	9	4	48	36	65
13	12	9	45	99	72
12	18	18	42	81	61
11	20	15	39	43	60
10	27	21	36	44	49
9	30	23	33	27	20
8	26	33	30	6	4
7	17	31	27	6	2
6	19	25	24	1	
5	17	21	21	1	1
4	13	13			
3	9	19			
2	7	5			
1	7	6			
0	5	9			
Total.........	262	263		429	432
Median.......	9.37	8.08		45.18	45.78
Average.......	8.77	7.61		44.01	44.84

constant error due to such lack of equivalence is furnished by the illustration given in the preceding paragraph. More exact evidence may be secured by arranging the duplicate forms in alternate order, and distributing them in this manner to pupils as they happen to be seated in the classroom. Thus, if Form 1 and Form 2 are being compared the first, third and fifth pupils will have a copy of Form 1, the second, fourth, sixth, etc., of Form 2. If it is decided to secure information for three forms of one test at a time, a similar arrangement will result in every third pupil having a copy of the same form. Form 2 and Form 3 of the Thorndike-McCall Reading Scale were arranged in this way and given to several hundred children. The same procedure was followed with reference to the Burgess Picture Supplement Scale for Measuring Silent Reading Ability. The distribution of scores from the different forms of the two tests is given in Table I. Both the median and the average scores for the Thorndike-McCall Reading Scale show that Form 2 and Form 3 are approximately equivalent. In the case of the Burgess Picture

Supplement Scale, the differences between the medians and the averages indicate a lack of equivalence which can not be neglected safely when precise comparisons are being made between scores yielded by the two forms.

By a similar method the equivalence of the duplicate forms of the Illinois General Intelligence Scale, Monroe's General Survey Scale in Arithmetic and Monroe's Standardized Silent Reading Tests, Revised, was studied.[6] The evidence collected for these three measuring instruments indicates that the different forms are slightly lacking in equivalence. This is especially true of the measures of rate yielded by the silent reading test. Thus, it has been considered necessary to give correction numbers whereby the scores yielded by one form of the test may be reduced to a basis comparable with those yielded by the other forms.

A similar study of the three forms of Monroe's Standardized Silent Reading Tests indicated a marked lack of equivalence.[7] In order to eliminate the constant error due to this cause corrections have been calculated which may be used to reduce the scores yielded by the different forms to a comparable basis. Separate sets of norms have been stated for each form.

3. **Evidence of constant errors due to instruction functioning as coaching.** When any considerable period of time elapses between two trials on a given test, the instruction which pupils receive during this interim may materially influence their second trial scores. In a recent investigation[8] by the Bureau of Educational Research it was found that for a group of 134 children the increase of the second trial scores on the Illinois General Intelligence Scale over the first trial scores was equivalent to slightly more than four years in mental age. The two trials were six months apart and hence the normal increase to be expected would be six months. If we assume that the first trial scores were accurate, it follows that the constant error introduced in the second trial scores was in the neighborhood of three

[6] Monroe, Walter S. "The Illinois Examination." University of Illinois Bulletin, Vol. 19, No. 9, Bureau of Educational Research Bulletin, No. 6. Urbana: University of Illinois, 1921, 70 p.

[7] Monroe, Walter S. "Report of Division of Educational Tests for '19-20." University of Illinois Bulletin, Vol. 18, No. 21, Bureau of Educational Research Bulletin No. 5. Urbana: University of Illinois, 1921, p. 19.

[8] Odell, Charles W. "The use of intelligence tests as a basis of school organization and instruction." University of Illinois Bulletin Vol. 20, No. 17, Bureau of Educational Research Bulletin, No. 12. Urbana: University of Illinois, 1922. 78 p.

and one-half years of mental age. Investigation revealed that the teachers of these pupils had given instruction which incidentally functioned as coaching and increased the scores on the second trial of the test.

In an unpublished study made by Mr. H. N. Glick, deliberate coaching on the Army Group Intelligence Scale Alpha was given to a number of pupils. The increases in the scores when the test was repeated amounted in some cases to several hundred percent of the original scores.

In dealing with measures of achievement it is more difficult to demonstrate the presence of constant errors. Instruction is expected to result in an increase in achievement. However, the use of a standardized educational test implies the existence of standard conditions with respect to recency of instruction. Furthermore, in many cases we are measuring merely a sample of a pupil's achievement. In such cases our measurements are valid only if this sample is representative. Hence the increase in the scores yielded by a second application of an achievement test may represent a combination of true growth and spurious growth. If the instruction has been very recent in the field of the test or if it has been concentrated upon the particular sample of achievement which the test measures directly the increase in scores will represent, for the most part, spurious growth. Additional evidence on this point will be given in the next section.

When two dimensions of a pupil's ability are measured separately as in the case of both rate and comprehension of silent reading, we find that frequently the magnitude of one dimension is increased at the expense of the other. This may be due to the instruction which pupils have received or to other directions given them at the time of taking the test. Unless the two dimensions are interpreted together the effect of such compensating relation will be similar to that of a constant error.

4. **Evidence of constant errors in measures of progress in educational experimentation.** When we attempt to secure a measure of progress in achievement in a school subject by taking the difference between the averages (or medians) of two sets of scores, we frequently find evidence that one or both sets of scores involves an unknown constant error. Table II gives certain gains in achievement which were obtained in an experiment to determine the relative effect

TABLE II. TWO SETS OF GAINS IN ACHIEVEMENT WHICH INDICATE THE PRESENCE OF CONSTANT ERRORS IN CERTAIN SETS OF SCORES, FIFTH GRADE

Group	No. of Pupils	Reading Rate		Reading Comprehension		Arithmetic	
		I	II	I	II	I	II
I	70	27.93	−15.78	.96	.35	23.82	21.45
II	72	3.67	22.11	1.21	1.86	14.72	5.44
III	326	− 4.77	33.25	.92	2.06	12.07	6.36
IV	133	− 6.60	22.90	.82	.95	17.04	10.09
V	157	9.29	27.35	1.48	2.12	10.65	5.83
VI	143	− 9.26	41.48	.08	2.36	4.69	5.38

of the number of sections into which a class was divided.[9] The six experimental groups were taught under the same conditions with the one exception regarding the number of sections into which the classes were divided. The tests used were Monroe's Standardized Silent Reading Test I, Revised, and Monroe's General Survey Scale in Arithmetic. Form 1 of these tests was given early in October, Form 2, the first of February, and Form 1 was repeated early the following May. The first gains were calculated by subtracting the average of the October scores from that of the February scores; the second, by subtracting the average of the February scores from that of the May scores. The two forms of these tests have been shown to be slightly lacking in equivalence, especially in the case of reading rate.[10] The gains in Table II, however, are evidence of the presence of constant errors in addition to those resulting from the slight non-equivalence of the different forms.

On the basis of our knowledge of the effect of practise we should expect the first gains to be larger than the second gains unless the variations of experimental conditions materially influenced the achievements of the pupils, which is extremely unlikely. We find in both reading rate and reading comprehension that the first gains are frequently less than the second. In three cases the first gain for reading rate is negative. In arithmetic the first gains are larger

[9] Monroe, Walter S. "Relation of sectioning a class to the effectiveness of instruction." University of Illinois Bulletin, Vol. 20, No. 11. Bureau of Educational Research Bulletin, No. 11. Urbana: University of Illinois, 1922. 18p.

[10] Monroe, Walter S. "The Illinois Examination," University of Illinois Bulletin, Vol. 19, No. 9, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921, p. 12-18.

than the second in all cases except one. The smaller gains during the first semester than the second and particularly the negative gains are evidence of the presence of a constant error in at least one of the sets of scores from which the gains were computed. The gain in reading rate shown for Group I is interesting; from October to February there is a very marked increase in rate; for the second semester the gain is negative. This suggests that the average February score was too large, i.e., it involved a positive constant error.

Similar evidence of the presence of constant errors in measures of achievement is found in a recent study of the relation of class size to school efficiency.[11] In this investigation, as in the one just described, experimental groups were arranged in pairs with the experimental conditions alternating in the two semesters. Especially in Grades V and VII the relative magnitude of gains made in the different semesters indicates the presence of a constant error in at least one set of scores from which the gains were computed.

In another investigation[12] conducted by the Bureau of Educational Research the average increases in mental age during a period of six months for two groups of children, each numbering about 3000, were found to be .4 years and .9 years. During the next six months for the same two groups the increases were 1.4 years and 1.0 years respectively. The normal increase in mental age during either of these intervals is of course six months. The obtained increase for the first period might be expected to be somewhat greater because of the presence of a constant error introduced by the general practise effect. However, in one case the difference between the first and second trial scores is less than six months and in both the increase is less than the corresponding differences between the second and third trial scores. No explanation was found for these inconsistent gains but they are evidence that in some way an unknown constant error was introduced in some if not in all of the scores. The facts of this illustration become even more striking when we note that the total of the two gains for the first group is 1.8 years and that for the second 1.9. Thus, when the total interval of twelve

[11] "Relation of size of class to school efficiency." University of Illinois Bulletin, Vol. 19, No. 45, Bureau of Educational Research Bulletin No. 10. Urbana: University of Illinois, 1922, p. 20.

[12] Odell, Charles W. "The use of intelligence tests as a basis of school organization and instruction." University of Illinois Bulletin, Vol. 20, No. 17, Bureau of Educational Research Bulletin No. 12. Urbana: University of Illinois, 1922. 78 p.

months is considered, the total increase in mental age is approximately the same for the two groups. On the other hand, if the two intervals of six months are taken, the increases in the mental age are radically different for the two groups.

In the same investigation if only the scores yielded by the Illinois General Intelligence Scale are considered, the gain between the first and second testings is 1.1 years. For the second period it is 1.4 years. A constant error due to practise effect is expected in these gains but it is surprising to find that the second gain, which is the difference between the second and third trial scores, involves the larger error.

In each of these illustrations we have evidence of the presence of a constant error for which the cause is obscure. Furthermore, the exact magnitude of the constant error is unknown. The obscurity of the cause is due in part to the large number of teachers and pupils participating in each of these educational experiments. The constant errors may have been due to changes in the interest and attitude of the teachers and pupils toward the test. However, it was not possible to secure any direct evidence on this point. The fact that the cause is obscure makes the possible presence of constant errors in such data a serious matter and tends to arouse suspicions regarding the accuracy of measurements of ability in large cooperative experiments.

5. **Evidence of constant errors in subjective scoring.** The evidence cited in the preceding pages has related to testing conditions. The scoring of the tests used was highly objective. In case the scoring of the test papers is not objective it is necessary to consider also the constant errors which may be introduced in this process. In the marking of examination papers and other pupil performances where the scorer is asked to exercise judgment, much evidence has been collected to show that two persons differ widely in the scores which they assign to the same pupil performances. These differences are due in part to the presence of a constant error resulting from the fact that one of the scorers tends to be more liberal than the other. In a recent investigation[13] several sets of pupil performances for which the scoring was rather highly subjective, were scored inde-

[13] Monroe, Walter S. "A critical study of certain silent reading tests." University of Illinois Bulletin, Vol. 19, No. 22. Bureau of Educational Research Bulletin No. 8. Urbana: University of Illinois, 1922. 52 p.

TABLE III. SUBJECTIVITY OF SCORING REPRODUCTIONS BY THE WORD-COUNTING METHOD

Test	Form	Grade	No. of scores	Scorers	Difference of average scores
Memory	I	IV	92	Y—C	− 9.9
Memory	I	IV	27	Y—K	− 5.1
Memory	II	IV	116	Y—C	− 2.0
Memory	I	VII	123	Y—K	− 7.5
Memory	II	VII	100	Y—C	− 8.2
Memory	II	VII	31	Y—K	+ 4.1
Reproduction	I	IV	94	L—K	+ 6.8
Reproduction	II	IV	31	L—C	— 1.6
Reproduction	II	IV	68	L—K	+ 4.7
Reproduction	I	VII	117	M—F	− 0.5
Reproduction	II	VII	113	F−C	− 6.0
Brown	I	IV	111	T—My	+12.8
Brown	II	IV	110	T—My	+ 6.9
Starch (No. 7)	I	VII	119	M—C	− 5.8
Starch (No. 6)	II	VII	121	M—C	− 2.0

pendently by two persons under the supervision of a third. A part of one table is reproduced from this report to furnish evidence of the presence of a constant error in the scores assigned by one or both of the scorers. The entries in the column headed "difference of average scores" were obtained by subtracting the average of the scores assigned by the second scorer from the average of those assigned by the first scorer. Some of these differences are relatively large. It appears that the scorer is not always consistent with respect to his constant error. Scorers Y and K show positive differences for one set of papers and negative differences for another set.

In the same investigation, eighty-six compositions were rated independently by two persons using the Willing Scale for Measuring Written Composition. The difference between the averages of their scores was 6.7.

Constant errors in first trial scores. As we have already indicated, first trial scores may involve constant errors. If there have been any departures from standard testing conditions we may expect to find the scores yielded too high or too low. It is possible to coach pupils for the first administration of a test as well as for later administrations. It may happen that where there has been no intentional coaching the instruction which they have received immediately prior to the taking of the test has served as preparation for the test.

Furthermore, if the norms are for pupils who are relatively unacquainted with testing procedure, test scores made by pupils who are accustomed to taking tests will involve a constant error with reference to these norms. At first the norms for our standardized educational tests were based upon scores obtained from pupils who had little or no experience in the taking of tests. This was necessarily so because such tests were new. As tests have become more widely used this factor of the testing conditions has changed, and it is probably true that norms for tests which have been recently standardized are based upon scores from many pupils who are familiar with general testing procedure. However, we have no specifications concerning the degree of acquaintance with the testing procedure for which the norms are stated.

In addition to the influence of instruction and acquaintance with testing procedures, constant errors may be introduced in first trial scores by the attitude of the pupils toward the test, by the way in which the test is explained to the pupils, and by a number of other factors which are subject to only partial control. In the case of handwriting the performances of pupils are very easily influenced by the type of directions given them. For example, in response to the instructions "Write as fast as you can" one college sophomore increased her rate of writing 77 letters per minute over her rate when writing for highest quality. One investigator[14] has presented evidence which shows that if pupils know they are being tested they will tend to write much more slowly than their normal rate. This reduction in rate is usually accompanied by an increase in quality. Similar results have been found for tests in other fields. The fact that test scores are influenced in this way by the directions given to pupils does not mean that they necessarily involve a constant error. It is only when these directions constitute departures from the standard testing conditions that we may expect constant errors. The evidence presented here merely shows what may happen when there are even slight departures from standard testing conditions.

Exact magnitude of constant errors can not be determined. In none of the cases cited to illustrate the presence of constant errors, was it possible to determine the exact magnitude of the constant error unless some basis for comparison was assumed. When a test

[14] Sackett, L. W. "Comparable measures of handwriting." School and Society, 4:640-45, October 21, 1916.

is repeated after a short interval of time the difference between the averages of the scores obtained from the two trials becomes the magnitude of the constant error in the second trial scores only if the first trial scores involve no constant error. Such an assumption may be justified in certain cases but one can never be certain that standard testing conditions prevailed in all details. Even when the examiner has exercised special care some of the more subtle factors of the testing conditions may not have been completely controlled. Unless good evidence can be produced in support of the assumption that the first trial scores involved a negligible constant error it is not safe to consider the difference between the averages of the two sets of scores as equivalent to the constant error. In more complex situations where a test is given three or more times for the purpose of measuring progress for two or more periods, it becomes more obvious that the exact magnitude of the constant error can not be determined. This condition has been indicated already in the evidence presented to show that constant errors were introduced in the data gathered in large cooperative educational experiments.

Altho one can not determine the exact magnitude of the constant error of measurement in a given case he can frequently present evidence to show that it probably does not exceed a certain amount. If his use of the data does not involve precise comparisons it may be possible to show that the constant error may be safely neglected. However, when precise comparisons are required and conclusions depend upon small differences between average or median scores the possible presence of constant errors makes such conclusions of doubtful validity.[15]

[15] In order to contrast the effects of the two types of errors upon the average and other derived measures, the consideration of the fourth question stated on page 7 with reference to constant errors is left until after the treatment of variable errors. The effect of both types of errors upon derived measures is considered in Chapter IV.

CHAPTER III

CAUSES, NATURE, AND MAGNITUDE OF VARIABLE ERRORS OF MEASUREMENT

1. **Evidence of variable errors of measurement secured when a test is repeated.** In order to secure evidence of the presence of variable errors of measurement it is necessary only to repeat a test after a short interval of time and compare the two scores of individual pupils. When this is done it is found that some pupils make a higher score on the first test and others on the second. In Table IV, two sets of scores yielded by the Monroe General Survey Scale in Arithmetic are given. The first pupil made a score of 51 on the first trial and 59 on the second. The difference in the two scores is — 8. Most of the differences are small. A few are relatively large. Approximately half are positive. The facts shown in this table are typical of the scores yielded by educational tests. For a few tests the scores involve smaller variable errors of measurement but for a number they are larger than in this illustration.

It should be noted that the differences between the two scores given in Table IV are not the variable errors of measurement. They are merely indicative of the presence of such errors and, in a crude way, of their magnitude. Neither set of scores can be considered true scores. Both are subject to variable errors and also, possibly, to an unknown constant error. In order to obtain the exact magnitude of the variable error of measurement for a given pupil it would be necessary to secure a true score and to subtract the obtained score from it. Such differences, when corrected for the constant error, would be the variable errors of measurement.

Method of describing the variable errors of measurement. As we have already indicated, it is impossible to determine the pupil's true score (see page 9). Furthermore, we always find variable errors and constant errors occuring in combination. It is, however, possible to secure a description of the magnitude of the variable errors of measurement which may be expected in the scores yielded by a given educational test when it is administered under standard testing conditions. Two sets of scores such as given in Table IV furnish the data upon which this description is based. These are obtained by

TABLE IV. SCORES YIELDED BY TWO APPLICATIONS OF MONROE'S GENERAL SURVEY SCALE IN ARITHMETIC TO A GROUP OF FIFTH GRADE PUPILS

Form I	Form II	Difference	Form I	Form II	Difference
51	59	−8	46	58	−12
49	60	−11	49	54	−5
46	60	−14	60	71	−11
77	84	−7	43	41	+2
42	43	−1	45	40	+5
43	43	0	30	24	+6
51	63	−12	28	23	+5
33	36	−3	46	38	+8
41	48	−7	34	32	+2
40	46	−6	59	56	+3
39	53	−14	63	72	−9
35	47	−12	42	45	−3
42	47	−5	45	53	−8
21	25	−4	48	73	−25
113	109	+4	51	59	−8
42	35	+7	53	56	−3
11	8	+3	68	68	0
23	21	+2	24	19	+5
45	38	+7	45	66	−21
54	49	+5	21	21	0
21	31	−10	37	36	+1
53	48	+5	38	39	−1
43	53	−10	39	50	−11
106	86	+20	50	64	−14
27	19	+8	30	43	−13
46	45	+1	33	42	−9
42	29	+13	65	74	−9
45	62	−17	69	86	−17
55	54	+1	51	59	−8
46	35	+11	43	43	0
38	32	+6	53	52	+1
17	15	+2	38	35	+3
53	61	−8	62	72	−10
52	65	−13	77	75	+2
50	58	−8	85	76	+9
41	43	−2	111	95	+16
37	48	−11	70	66	+4
26	34	−8	107	84	+23
57	65	−8	9	8	+1
51	59	−8	27	25	+2
34	41	−7	39	32	+7
26	36	−10	25	23	+2
22	24	−2	59	55	+4
59	64	−5	57	47	+10
49	37	+12	41	36	+5
61	75	−14	56	62	−6

two applications of the same test or of duplicate forms of a test to a group of representative pupils. These two applications should be separated by only a small interval of time. One type of description of the magnitude of the variable error of measurement is obtained by calculating the coefficient of correlation between these two sets of scores. This, when applied to an educational test, is called the coefficient of reliability. It indicates in a rough way the magnitude of the variable error of measurement, and is unaffected by the presence of any constant error of measurement in the two sets of scores from which it is calculated.

The coefficient of reliability an unsatisfactory description of variable errors of measurement. Altho the coefficient of reliability is an index of the variable error of measurement, a given coefficient, say .85, can not be interpreted directly in terms of the magnitude of these variable errors of measurement. Experienced persons are able to attach a reasonably concrete meaning to a given reliability coefficient but to an inexperienced person a coefficient of reliability, say .72 or .95, can have little more than a very general meaning.

Under certain conditions reliability coefficients furnish us with an index of the relative magnitude of the variable errors to be expected in the scores yielded by different tests. For example, if the reliability coefficient for one test is .65 and for another, .85, we should expect to find the variable errors of measurement for the second test much smaller. However, considerable caution must be exercised in comparing coefficients of reliability. The writer has shown[1] that the correlation of the two scores yielded by a given test is much smaller when the scores are taken from a single grade than when taken from a sequence of two or more grades. In one illustration when the scores were assembled separately for half-grade groups the highest coefficient of correlation was .57. For one half grade it was .12 and for another .27. When the scores for all grades from III-B to VIII-A, inclusive, were assembled the coefficient of correlation was .76. In this illustration there were from two hundred to four hundred cases in each half grade. Hence the variations can not be explained on the basis of sampling.

Probable error of measurement used to describe the variable error. Altho the coefficient of reliability is unsatisfactory it may

[1] Monroe, Walter S. An Introduction to the Theory of Educational Measurements. Boston: Houghton Mifflin Company, 1923, p. 356.

be used as a basis for calculating the probable error of measurement. The formula[2] for this is given below.

$$P.E._M = .6745 \frac{\sigma_1 + \sigma_2}{2} \sqrt{1 - r_{12}}$$

In this formula r_{12} is the coefficient of correlation between first and second trial scores. σ_1 is the standard deviation of the distribution of the first trial scores and σ_2 a corresponding measure for the second trial scores.

It should be noted that the probable error of measurement does not give the magnitude of the variable error of measurement for any one pupil. It gives merely the limits between which we may expect to find 50 percent of the variable errors of measurement of a given group of scores. For example, the probable error of measurement for the rate score yielded by the Courtis Silent Reading Test No. 2 has been found to be 19.3.[3] This means that 50 percent of the variable errors of measurement were greater than 19.3, approximately one-half of them being positive. It also means that 50 percent of them were not larger than 19.3 nor smaller than — 19.3. In the case of a given pupil we can state only the chances that the probable variable error of measurement of his score does not exceed certain limits; as, for example, in the Courtis Silent Reading Test referred to, the chances were just even that the variable error of measurement in his score was not larger than ± 19.3. The chances are 4.6 to 1 that the variable error of measurement of his score is between — 38.6 and +38.6. The chances for other limits also may be stated.

The magnitude of variable errors to be expected in educational measurements. In the data given to show the presence of both constant and variable errors in our educational measurements their magnitude has been indicated. It is clear that they are much greater than the corresponding errors in physical measurements. In another place the writer has discussed the relative magnitude of the errors in the scores yielded by standardized tests and the errors in the

[2] For an explanation of this formula see, Monroe, Walter S. An Introduction to the Theory of Educational Measurements. Boston: Houghton Mifflin Company, 1923, p. 347-56.

[3] Monroe, Walter S. "A critical study of certain silent reading tests." University of Illinois Bulletin, Vol. 19, No. 22, Bureau of Educational Research Bulletin No. 8. Urbana: University of Illinois, 1922, p. 33.

grades assigned to examination papers.[4] The evidence presented indicates that the variable errors of measurement for a number of widely used standardized educational tests are only slightly less than the variable errors of measurement for written examinations. Some additional data with reference to the variable errors of the scores yielded by standardized tests will be helpful in arriving at the true understanding of their magnitude.

In a critical study of a group of silent reading tests[5] it was shown that the probable error of measurement for some tests was greater than 25 percent of the average score. In fact, for Brown's Silent Reading Test it was found to be more than 50 percent. In the tests which make up the Illinois Examination, only twelve of forty-two probable errors of measurement which were calculated were greater than 10 percent of the average score.[6] The authors of the Stanford Achievement Test[7] announce that the probable error of measurement for this battery of tests is approximately two months of educational achievement. The coefficients of reliability are high. It is likely that these authors have succeeded in reducing the variable errors of measurement to a lower minimum than has been secured by others. This has been accomplished in part through extending the length of the test.

Using scores which were the medians of eight independent ratings of English compositions by means of the Nassau County Supplement to the Hillegas Scale, Hudelson[8] has given coefficients of reliability ranging from .69 to .84. The writer has estimated that if the ratings of a single judge had been used the coefficients of correlation would have been in the neighborhood of .40 instead of ranging

[4] Monroe, Walter S., and Souders, L. B. "The present status of written examinations." University of Illinois Bulletin, Vol. XXI, No. 13. Bureau of Educational Research Bulletin No. 17. Urbana: University of Illinois.

[5] Monroe, Walter S. "A critical study of certain silent reading tests." University of Illinois Bulletin, Vol. 19, No. 22, Bureau of Educational Research Bulletin No. 8. Urbana: University of Illinois, 1922, p. 52.

[6] Monroe, Walter S. "The Illinois Examination." University of Illinois Bulletin, Vol. 19, No. 9, Bureau of Educational Research Bulletin No. 6. Urbana: University of Illinois, 1921, p. 49.

[7] Kelley, T. L., Ruch, G. M., and Terman, L. M. Stanford Achievement Test Manual of Directions. Yonkers: World Book Company, 1923.

[8] Hudelson, Earl. "English composition, its aims, methods, and measurement." Twenty-second Yearbook of the National Society for the Study of Education, Part I. Bloomington, Illinois: Public School Publishing Company, 1923, p. 62.

from .69 to .84, and the probable error of measurement would have been a little more than six-tenths of one step of the scale used. This may appear relatively small but when we examine the norms we find that the unit used is relatively large. The average increase in norms from the fourth to twelfth grades, inclusive, is only slightly more than four-tenths of a unit per year. Between the eighth and ninth grades the increase is only two-tenths of a unit. The greatest yearly increase is six-tenths of a unit. Thus we have here a probable variable error of measurement which is relatively large.

CHAPTER IV

THE EFFECT OF CONSTANT AND VARIABLE ERRORS UPON DERIVED MEASURES

The effect of errors of measurement upon derived measures. There seems to be a prevailing idea that the effect of errors of measurement upon such derived measures as the average, median, standard deviation and coefficient of correlation, may be safely neglected if the derived measure is based upon a sufficiently large number of cases. This is only partially true. A constant error in the original data makes the average in error by the amount of the constant error. Any increase in the number of cases has no effect upon the magnitude of the constant error. It can not be eliminated or even reduced unless we are able to determine its magnitude, in which case it may be subtracted from the average. The same situation prevails for the median. However, a constant error does not affect the standard deviation and other measures of variability. Neither does it affect the coefficient of correlation.

As we have already pointed out in the preceding pages, we are seldom able to determine even approximately the magnitude of the constant errors of the scores yielded by educational tests. We have evidence only of their presence. Hence, it is impossible to make any accurate correction for a constant error. It has been estimated that second trial scores are about 10 percent larger than first trial scores but studies of different tests indicate that this constant error of measurement varies widely. For some tests the difference between first trial scores and second trial scores is much less than for others. It is also doubtless less for some groups of pupils than for others. When pupils are acquainted with the testing procedure the increase of second trial scores over first trial scores may be very slight, especially if the children are acquainted with other tests having similar structure. When compared with first trial scores, third trial scores involve a somewhat larger constant error than those obtained from the second trial, and beyond the third trial it is likely that there is some increase.

Unlike constant errors of measurement variable errors tend to neutralize each other in the average. The reason for this is easily understood because approximately one-half of the variable errors of measurement are negative and the other half positive. If we increase the number of cases the magnitude of the variable errors of measurement in the average is decreased. The relation is given by the following formula:

$$\text{P.E.}_{\text{M}} \text{ average} = \frac{\text{P.E.}_{\text{M}}}{\sqrt{\text{N}}}$$

It should be noted that the error of the average due to the presence of variable errors of measurement in the data can not be explicitly defined. It is necessary to describe it in terms of the probable error (P.E._{M} average). The above formula gives the limits between which the chances are even that the error of the average will fall.

Variable errors of measurement tend to make the standard deviation and other measures of variability larger than they would be otherwise. The relation between the obtained standard deviation and the true standard deviation is given by the following formula:

$$\sigma_{\text{true}} = \sigma_{\text{obtained}} \sqrt{r_{12}}$$

In this formula r_{12} is the coefficient of reliability of the scores concerned. Since N does not appear in this formula it follows that increasing the number of cases does not have any effect upon the variable error of measurement of a measure of variability.

The presence of variable errors of measurement in our data always tends to decrease the coefficient of correlation.[1] If each of the two sets of facts whose relationship we are studying has been measured in duplicate, it is possible to correct for the effect of these variable errors of measurement. For example, if it is desired to secure the true correlation between ability to reproduce a selection read and ability to answer questions upon it, we may secure a corrected coefficient of correlation by measuring each of these abilities twice. One formula which has been used for this purpose is the following:

$$r_{pq} = \frac{\sqrt{(r_{p_1q_2})\ (r_{p_2q_1})}}{\sqrt{(r_{p_1p_2})\ (r_{q_1q_2})}}$$

r_{pq} here indicates the true correlation between two series of measures, p and q, of the facts A and B.

p_1 and p_2 are two independent measures of A.

q_1 and q_2 are two independent measures of B.

[1] Thorndike, E. L. Theory of Social Measurements. New York: Teachers College, Columbia University, 1916, second edition, p. 178.

$r_{p_1q_2}$ is the correlation obtained from the first measure of A and the second measure of B.

$r_{p_2q_1}$ is the correlation obtained from the second measure of A and the first measure of B.

$r_{p_1p_2}$ is the correlation between the two measures of A.

$r_{q_1q_2}$ is the correlation between the two measures of B.

In a recent study[2] it was desired to secure the correlation between the comprehension in silent reading as measured by Monroe's Standardized Silent Reading Tests with the scores yielded by a test of memory. It was recognized that both sets of scores involved variable errors of measurement which would materially decrease the magnitude of the coefficient of correlation. For this reason it was arranged to measure each trait in duplicate. The coefficients of correlation obtained from correlating each measure of comprehension yielded by a silent reading test with the two measures of memory were .31, .33, .31, and .35. The coefficient of reliability for the memory test was .35 and for the silent reading test, .73. In this illustration $r_{p_1p_2}$ equals .65, $r_{q_1q_2}$ equals .35, $r_{p_1q_2}$ equals .33, and $r_{p_2q_1}$ equals .31. Substituting these values in the formula given above we obtain for the corrected coefficient of correlation between comprehension and memory a value of .67. This gives some indication of the effect of the variable errors of measurement in these two sets of scores upon the coefficient of correlation between comprehension and memory.

[2] Monroe, Walter S. "A critical study of certain silent reading tests." University of Illinois Bulletin, Vol. 19, No. 22, Bureau of Educational Research Bulletin, No. 8. Urbana: University of Illinois, 1922, p. 41.

CHAPTER V

EFFECT OF ERRORS UPON USE OF EDUCATIONAL MEASUREMENTS

The attitude toward educational measurements as affected by the recognition of errors. The effective use of any instrument depends upon a frank recognition of its limitations. Standardized educational tests are no exception. They have been advertised by their authors and by others as measuring instruments which are distinctly superior to ordinary written examinations. At first attention was centered upon their use and few studies were made of the errors involved in the scores yielded. It has been apparent for some time, however, that test scores are subject to errors which sometimes are astonishingly large. This bulletin has been written to call attention to the nature of these errors, their magnitude, and their effect upon the average and other derived measures.

Those of us who have not been concerned with the errors involved in educational measurements may tend to feel doubtful of the value of standardized tests after realizing the nature and magnitude of the errors encountered. In the judgment of the writer, this effect should not be produced. We cannot expect to make our use of educational tests most effective until we are informed concerning the limitations of the measures yielded. A frank recognition of the presence of both constant and variable errors should enable the users of educational tests to do their work more efficiently. In many cases they can avoid erroneous interpretations which they otherwise would make. Furthermore, it is only by understanding the nature of the errors which are likely to be encountered that we can take steps to reduce them to the lowest possible minimum, and be aided in the construction of improved measuring instruments because of our knowledge of the defects of the present ones.

The writer does not advise the discontinuance of the use of standardized educational tests because the measures yielded have been shown to involve both variable and constant errors larger than many of us supposed. There is abundant evidence to show that the use of educational tests in our schools is increasing their efficiency. Still greater improvement may be expected when measuring instruments are used more intelligently.